

Improved variance estimation along sample eigenvectors

Anne Hendrikse Raymond Veldhuis Luuk Spreeuwers
University of Twente
Fac. EEMCS, Signals and Systems Group
P.O.Box 217, 7500 AE Enschede, The Netherlands
a.j.hendrikse@utwente.nl

Abstract

Second order statistics estimates in the form of sample eigenvalues and sample eigenvectors give a sub optimal description of the population density. So far only attempts have been made to reduce the bias in the sample eigenvalues. However, because the sample eigenvectors differ from the population eigenvectors as well, the population eigenvalues are biased estimates of the variances along the sample eigenvectors. Therefore correction of the sample eigenvalues towards the population eigenvalues is not sufficient. The experiments in this paper show that replacing the sample eigenvalues with the variances along the sample eigenvectors often results in better estimates of the population density than replacing the sample eigenvalues with the population eigenvalues.

1 Introduction

An important aspect in statistic learning is the modeling of data distributions. A common assumption is that the distribution of the data after some preprocessing can be characterised by the second order statistics. The second order statistics of a multidimensional distribution are described by a covariance matrix, which we denote the population covariance matrix. The population covariance matrix is given by $\Sigma = \mathcal{E}(\tilde{\mathbf{x}} \cdot \tilde{\mathbf{x}}^T)$, with $\tilde{\mathbf{x}} = \mathbf{x} - \mathcal{E}(\mathbf{x})$, where $\mathcal{E}()$ is the expectation operator and \mathbf{x} is a random variable representing the data generating process.

The covariance matrix can be decomposed such that $\Sigma = \mathbf{E} \cdot \mathbf{D} \cdot \mathbf{E}^T$, where \mathbf{E} is a rotation matrix. Each column of \mathbf{E} is an eigenvector. Diagonal matrix \mathbf{D} has the eigenvalues of Σ at the diagonal. The i^{th} eigenvector points in the direction with the largest variance after the subspace formed by the first $i - 1$ eigenvectors has been removed. The i^{th} eigenvalue gives the variance in this direction. As an example, the black curve in figure 1 shows the variance of a two dimensional, zero mean distribution with eigenvalues 2 and 1. The first eigenvector is along the vertical axis and the second eigenvector is along the horizontal axis.

Usually, the population covariance matrix is unknown beforehand and needs to be estimated from a set of examples, the training set. A commonly used estimator is the sample covariance matrix: $\hat{\Sigma} = \frac{1}{N-1} \mathbf{X} \cdot \mathbf{X}^T$, where each of the N columns of \mathbf{X} consists of a sample from the training set with the mean of the set subtracted. The decomposition results of the sample covariance matrix are denoted by sample eigenvectors and sample eigenvalues. As a result of the limited number of examples, the estimates contain errors. For the sample covariance matrix, these errors have a zero mean: the estimator is unbiased. However, the sample eigenvalues are a biased estimate of the population eigenvalues. We have shown previously that reduction of

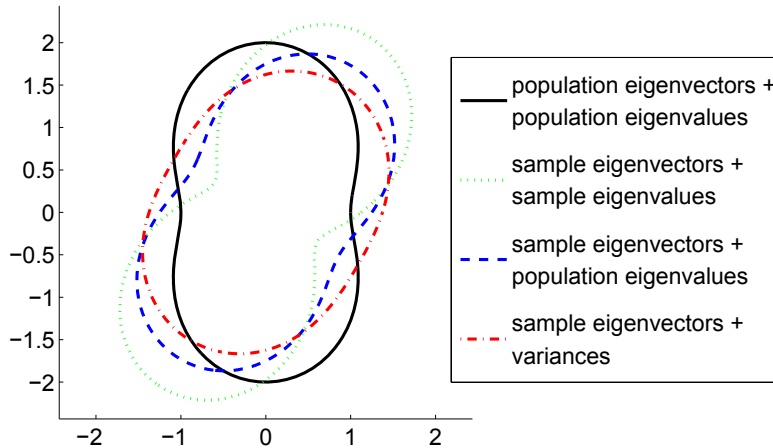


Figure 1: Variance curves of approximations of an example distribution.

this bias is possible [4]. In this paper we argue that even if the entire bias in the sample eigenvalues were removed, the combination of corrected sample eigenvalues and uncorrected sample eigenvectors still provides a sub optimal description of the population distribution. We present a study on another solution to correct the sample eigenvalues. The results of the study cannot be used directly in practice because we make extensive use of the population eigenvectors and these are usually unknown in estimation problems. A practical approach is shown in our work in [4]. Because our interest lies in the area of biometrics, some of the experiments use data models used in biometrics.

The remainder of the paper is outlined as follows: in the next section we present some analysis on the sample eigenvalues and eigenvectors, showing that besides the bias in the eigenvalues, the inner product of the sample eigenvectors with the population eigenvalues also shows some sort of bias. This leads to a new method of replacing the sample eigenvalues: the variance estimate. In section 3 we compare replacement of the sample eigenvalues with the population eigenvalues and replacement with variances experimentally, using the Kullback Leibler divergence as an evaluation criterion in section 3.1 and verification scores in section 3.2. We draw conclusions in section 4.

2 Eigenvector and eigenvalue analysis

2.1 Eigenvalue analysis

Estimators attempt to obtain parameters from a limited number of samples. As a result, the estimate contains errors. To describe these errors, often Large Sample Analysis (LSA) is performed on the estimator. In LSA the limit behaviour of the error of the estimator is described under the assumption that the number of samples grows to infinity. In LSA, the sample eigenvalues show no bias. However, in biometrics the number of samples (N) is limited and in the same order as the number of dimensions (p) or even lower. Therefore LSA should not be applied.

In General Statistical Analysis (GSA) another limit is considered: $N, p \rightarrow \infty$ while $\frac{p}{N} \rightarrow \gamma$, where γ is some positive constant. In this limit, the sample eigenvalues do show a bias and a relation between the sample eigenvalues and the population eigenvalues is given for a large class of data distributions by the Marčenko Pastur equation [7]. The bias is such that the largest sample eigenvalue is larger than the largest population eigenvalue and the smallest sample eigenvalue is smaller than the

smallest population eigenvalue. In figure 1 the dotted line gives the variance estimates if the sample eigenvalues are biased estimates and the sample eigenvectors are off as well.

In the remainder of the article we assume that the bias can be fully compensated for and the remaining fluctuations in the sample eigenvalues are minimal, allowing perfect estimation of the population eigenvalues. However, combining the population eigenvalues with the sample eigenvectors still provides an estimate of the distribution containing significant deviations caused by deviations between sample and population eigenvectors. In the example this is shown by the dashed curve in figure 1.

2.2 General Statistical Analysis of the sample eigenvectors

We expect that the sample eigenvectors are generally off compared to the population eigenvectors as well when studied in GSA. To our knowledge there is no theory available which describes the relation between sample eigenvectors and population eigenvectors similar to the Marenko Pastur equation, but Anderson did derive an expression for the distribution of the sample eigenvectors if the data is Gaussian [1]. Mestre presented some results on the estimations of subspaces belonging to the eigenvalues in the GSA limit in [6]. In [2] the inner product between the eigenmatrix and a random unitary vector is considered. Our approach differs because we do not consider a random vector, but the population eigenvectors.

To demonstrate that the sample eigenvectors are generally off compared to the population eigenvectors, we did a synthetic eigenvector estimation experiment. In the experiment, synthetic Gaussian data is generated with all eigenvalues chosen uniformly between 0 and 1. The sample covariance matrix is determined, which is decomposed to find the sample eigenvectors. We then determined the component of a sample eigenvector $\hat{\mathbf{E}}_{:,m}$ in the subspace spanned by the K population eigenvectors $\{\mathbf{E}_{:,k} | k = 1 \dots K\}$ with the smallest population eigenvalues, which is given by $\sqrt{\sum_{k=1}^K (\hat{\mathbf{E}}_{:,m}^T \mathbf{E}_{:,k})^2}$.

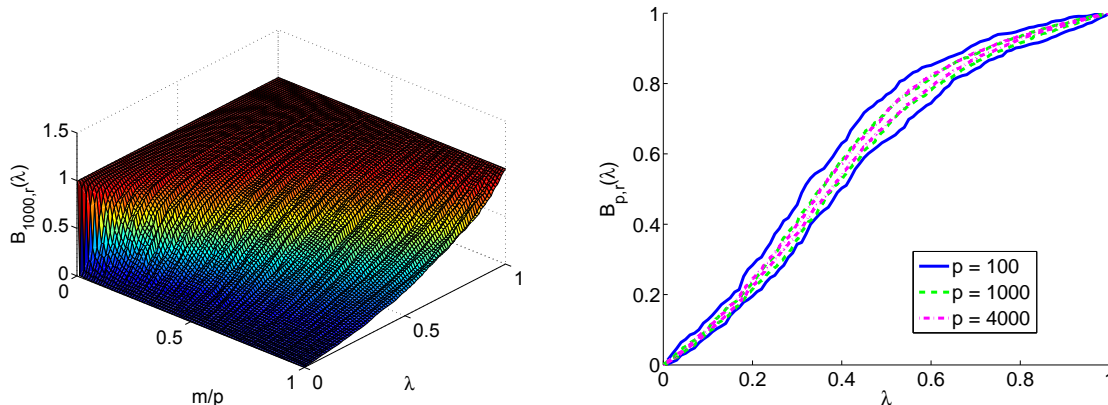
If we repeat this procedure for $K = 0 \dots p$ we get a function which starts at 0 and increases to 1 for $K = p$. If we consider this component as a function of the largest population eigenvalue instead of the largest index, we can describe the components as a distribution function $B_{p,m}(\lambda)$ given by:

$$B_{p,m}(\lambda) = \sqrt{\sum_{k=1}^p (\hat{\mathbf{E}}_{:,m}^T \mathbf{E}_{:,k})^2} u(\lambda - \lambda_k) \quad (1)$$

where $u(\cdot)$ is the step function and λ_k is the k^{th} element of a vector with the population eigenvalues sorted from small to large. We hypothesize that under the GSA limit, $B_{p,m}(\lambda)$ like the eigenvalue distributions converges to a fixed distribution function.

Figure 2a shows the distribution functions for all sample eigenvectors. In the experiment $p = 1000$ and $\gamma = \frac{1}{2}$. The sample eigenvalue index is scaled with p so it ranges from 0 to 1. As can be seen from the figure, the smallest sample eigenvectors are confined to the space of the smallest population eigenvectors. The largest sample eigenvectors have a component in almost all population eigenvectors.

Figure 2b shows why we hypothesize that $B_{p,m}(\lambda)$ converges in the GSA limit, similar to the eigenvalues. In the figure, we show the mean plus and minus a standard deviation of the middle sample eigenvector for three different configurations: $p = 100$, $p = 1000$ and $p = 4000$, all with $\gamma = \frac{1}{2}$, with the number of repetitions equal to 20, 8 and 4 respectively. The mean curves were all almost indistinguishable, while the standard deviation curves tighten around the mean. It seems that the distribution converges to a fixed distribution in the GSA limit.



(a) Distribution of the component of the sample eigenvectors in the population eigenvector subspace.

(b) Convergence example of $B_{p,r}(\lambda)$. The curves show the mean plus and minus one standard deviation of $B_{p,r}(\lambda)$ based on a number of repetitions per configuration.

Figure 2: Inner product example of sample eigenvector variation

Because the sample eigenvector is not in the same direction as the population eigenvector, the corresponding population eigenvalue will in general be an erroneous estimate of the variance along the sample eigenvector. We therefore propose to replace the sample eigenvalues with the variance along the sample eigenvectors (\mathbf{v}_m for sample eigenvector $\hat{\mathbf{E}}_{:,m}$):

$$\mathbf{v}_m = \sum_{k=1}^p \left(\left(\hat{\mathbf{E}}_{:,m}^T \cdot \mathbf{E}_{:,k} \right)^2 \lambda_k \right) \quad (2)$$

In the example in figure 1 this correction leads to the dash-dotted curve. Since this is still not a perfect match with the real distribution, the question is which of the two sample eigenvalue replacements is better: population eigenvalues or variances.

3 Experimental comparison between population eigenvalue substitution and variance substitution

To compare sample eigenvalue replacement by population eigenvalues with replacement by variances a measure is required which describes how close a replacement distribution is to the population distribution. We chose two measures: the Kullback Leibler divergence and verification rates. With both measures we did an experiment on synthetic data.

3.1 Comparison based on the Kullback Leibler divergence

The first experiment compares the replacement with population eigenvalues and the replacement with variances by determining the Kullback Leibler divergence[5] between the replacement distributions and the population distribution. For the experiment we generate Gaussian distributed samples. The Kullback Leibler divergence between two Gaussian distributions N_0 and N_1 with covariance matrices Σ_0 and Σ_1 respectively

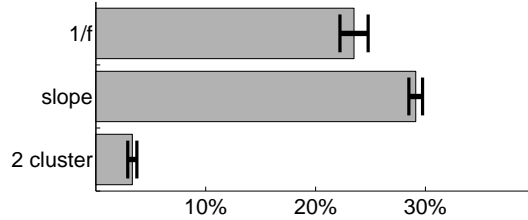


Figure 3: Kullback Leibler divergence reduction when replacing sample eigenvalues with variances instead of population eigenvalues. The gray bars indicate the average improvement, while the thick black lines indicate the standard deviation.

and equal mean is given by [8]:

$$d_{\text{KL}}(N_0|N_1) = \frac{1}{2} \left(\log \left(\frac{\det \Sigma_1}{\det \Sigma_0} \right) + \text{tr}(\Sigma_1^{-1} \Sigma_0) - p \right) \quad (3)$$

In the experiment, N_1 was always the population distribution and N_0 was the replacement distribution N_{eig} for replacement with population eigenvalues and N_{var} for replacement with variances.

We used three different population eigenvalue distributions:

- 2 cluster: half of the population eigenvalues have a value of 2 and the other half have a value of 1.
- slope: the population eigenvalues are distributed uniformly between 1 and 2.
- 1/f: the population eigenvalues are set to $1/f$, where f is the index of the eigenvalue.

In all configurations we generated 500 samples with a dimensionality of 100. The first two configurations are therefore similar to the experiments by Karoui [3]. The 1/f is an eigenvalue model often used in biometrics. We repeated the experiments 100 times for each configuration.

We determined the divergence improvement between the eigenvalue replacement and the variance replacement by

$$s(N_{\text{eig}}, N_{\text{var}}) = (d_{\text{KL}}(N_{\text{eig}}|N_1) - d_{\text{KL}}(N_{\text{var}}|N_1)) / d_{\text{KL}}(N_{\text{eig}}|N_1) \quad (4)$$

Figure 3 shows the average and standard deviations of $s(N_{\text{eig}}, N_{\text{var}})$. In both the 2 cluster and the slope configuration the replacement with variances gives a considerable improvement of the density estimation. Even though the improvement in the 1/n configuration is not as large, using variances instead of population eigenvalues is still better.

3.2 Comparison based on verification experiments

In the second experiment we performed a verification experiment with synthetic data. Repeating the experiment for both the replacement with population eigenvalues and the replacement with variances gives verification rates for both methods. The objective is to determine which method gives the best verification performance.

We designed a verification experiment based on the Linear Discrimination Analysis (LDA) model, in which all classes are distributed with a normal distribution which only has a different mean per class. We generated a training set with 150 classes with 2 samples per class. The class means (μ_c) have a normal distribution with zero mean

and covariance matrix Σ_b , the samples within class c are distributed $N(\boldsymbol{\mu}_c, \Sigma_w)$, a normal distribution with mean $\boldsymbol{\mu}_c$ and covariance matrix Σ_w .

For testing we generated a set of probe samples distributed with the same distribution parameters as used for generating the training set, containing 400 classes and 2 samples per class. For enrollment we used the real class means instead of estimating it from a number of samples. All data sets had a dimensionality of 100.

Since LDA and classification is only influenced by the ratio between the within and the between class distributions, one of the distributions can be fixed. We therefore chose to fix Σ_b at $\frac{1}{10}\mathbf{I}$, where \mathbf{I} is an identity matrix. For Σ_w we can leave the eigenvector matrix also at \mathbf{I} and only have to chose an eigenvalue distribution. We tested four eigenvalue distributions:

- 2 cluster: half of the eigenvalues equal to 0.5 and the other half equal to 1.
- slope distribution: all eigenvalues chosen uniformly between 0.5 and 1.
- 1/f: the f^{th} eigenvalue is set to $30/f$.
- toeplitz: the eigenvalues of the matrix with elements $0.5^{|i-j|}$ where i and j are the row and column indices respectively.

During training, instead of using averages of the class samples, we used the real class means as estimates, so that the estimate of the between class covariance matrix does not contain crosstalk of the within class covariance matrix caused by fluctuations in the class mean estimates. After estimating the within class and between class covariance matrices, we performed four corrections: in the first correction, we kept the sample eigenvalues. In the second correction, we replaced the sample eigenvalues with the population eigenvalues. In the third correction, we replaced the sample eigenvalues with variances. In the last "correction", we used the population eigenvectors and population eigenvalues to get the results in case of correct estimation.

In the verification experiments, we used the log likelihood ratio (equation 5) to determine whether a sample originated from the claimed class.

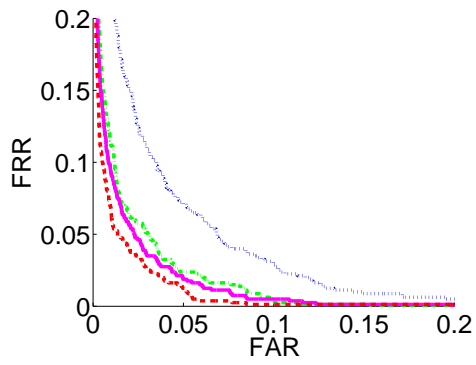
$$L(\mathbf{x}, c) = -(\mathbf{x} - \boldsymbol{\mu}_c)^T \Sigma_w^{-1} (\mathbf{x} - \boldsymbol{\mu}_c) + (\mathbf{x} - \boldsymbol{\mu}_t)^T \Sigma_t^{-1} (\mathbf{x} - \boldsymbol{\mu}_t) \quad (5)$$

If the likelihood ratio for a sample \mathbf{x} and class c combination is above a threshold, the sample is considered to belong to class c . By changing the threshold a trade-off can be made between the ratio of samples incorrectly accepted to a class (False Accept Rate, FAR) and the ratio of samples erroneously rejected from a class (False Reject Rate, FRR).

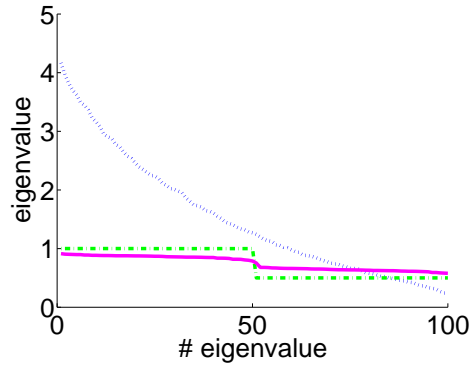
In figure 4 we show the results of the verification experiments. The curves on the left show Detection Error Trade off (DET) curves, which plot FRR against the corresponding FAR. In the right column we show the within class eigenvalues for the different corrections. The between class eigenvalues are the same for the eigenvalue correction and the variance correction, namely all $\frac{1}{10}$. In the left column we also show the classification results when the population parameters are used, so the population eigenvalues as well as the population eigenvectors.

In the DET curves there is a large reduction between no correction and eigenvalue correction in all configurations. The difference between the eigenvalue correction and the variance correction is much smaller, but as the population correction shows, the eigenvalue correction is already much closer to the best estimate than the original sample estimate. The variance correction gives a significant improvement in both the 2 cluster configuration and the slope configuration (also if the experiment is repeated a number of times). The Toeplitz results also seem to have improved with the variance correction, however the difference is too small to be significant.

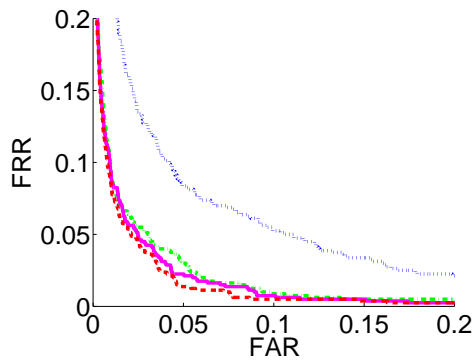
The scree plots on the right show a general trend we expected: the variance correction reduces the largest eigenvalues while increasing the smallest eigenvalues. With the 1/f configuration however the modification is small. With the Toeplitz, the modification is considerable even though it has only a small effect on the verification rates.



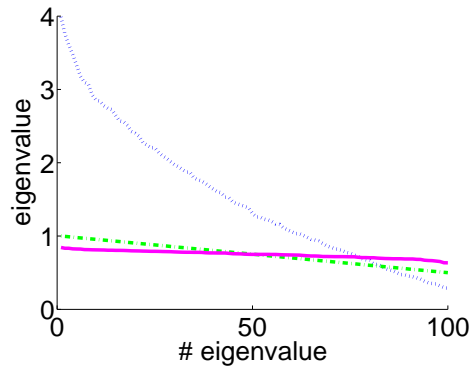
(a) 2 cluster verification rates



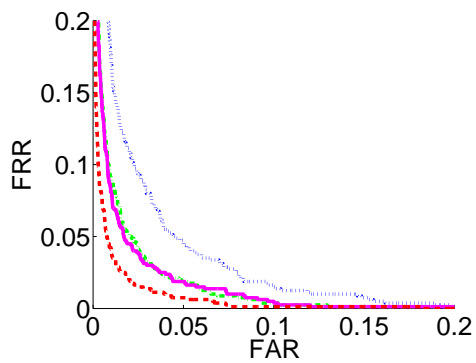
(b) 2 cluster within eigenvalue scree plots



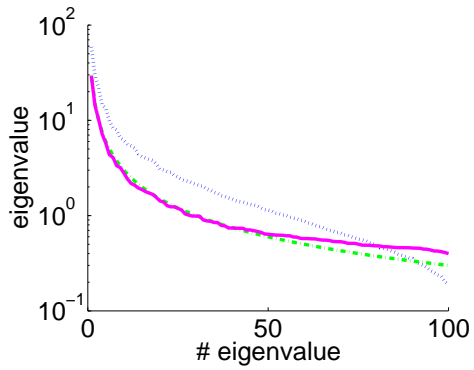
(c) Slope verification rates



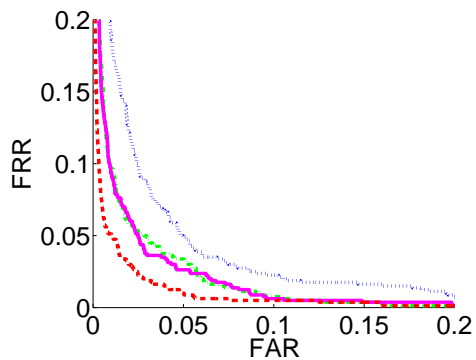
(d) Slope within eigenvalue scree plots



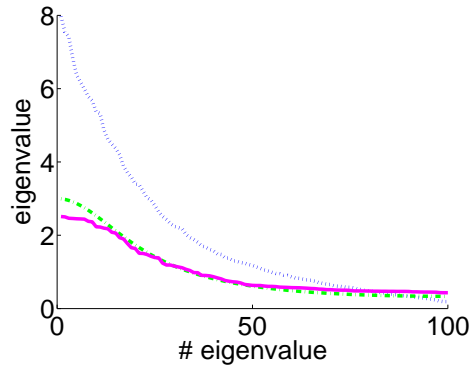
(e) 1/f verification rates



(f) 1/f within eigenvalue scree plots



(g) Toeplitz verification rates



(h) Toeplitz within eigenvalue scree plots

Figure 4: Verification experiment results. The curves are the results as follows: dotted = sample eigenvectors with sample eigenvalues, dash dotted = sample eigenvectors with population eigenvalues, solid = sample eigenvectors with variances and dashed = population eigenvectors with population eigenvalues.

4 Conclusion & discussion

The combination of population eigenvalues with sample eigenvectors does not lead to an optimal description of the population distribution. We showed that the sample eigenvectors will be off compared to the population eigenvectors, which makes the population eigenvalues a biased variance estimate along the sample eigenvectors. For example, if there is just one population eigenvalue with the maximum value, then none of the sample eigenvectors will be along the corresponding population eigenvector and therefore the maximum variance along a sample eigenvector will be smaller than the maximum population eigenvalue.

We suggested to replace the sample eigenvalues with the variances along the sample eigenvectors. In the experiments, using the variances instead of the population eigenvalues showed as good as and often better estimates of the population distribution. The 2 subset method in [4] already showed a practical implementation of using variances.

References

- [1] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley series in probability and mathematical statistics. John Wiley & Sons, 2 edition, 1984.
- [2] Z. D. Bai. Methodologies in spectral analysis of large dimensional random matrices, a review. In *Statistica Sinica*, number 9, pages 611–677, 1999.
- [3] N. El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *ArXiv Mathematics e-prints*, september 2006.
- [4] A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis. Eigenvalue correction results in face recognition. In *Twenty-ninth Symposium on Information Theory in the Benelux*, pages 27–35, 2008.
- [5] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86, 1951.
- [6] X. Mestre. Estimating the eigenvalues and associated subspaces of correlation matrices from a small number of observations. In *Proc. of the 2nd Int. Symposium on Communications, Control and Signal Processing*, Marrakech (Morocco), 2006.
- [7] Jack W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *J. Multivar. Anal.*, 55(2):331–339, 1995.
- [8] M. Tumminello, F. Lillo, and R. N. Mantegna. Kullback-leibler distance as a measure of the information filtered from multivariate data. *Physical Review E*, 76:031123, 2007.